

Statistics 210
Statistical Methods

Hal S. Stern
Department of Statistics
University of California, Irvine
sternh@uci.edu

- Dictionary definitions:
 - statistic - a single term or datum; a quantity that is computed from a sample
 - statistics - a branch of mathematics dealing with the collection, analysis, interpretation and presentation of masses of numerical data
- Points of emphasis
 - numbers with a context
 - inference from sample to population
 - interaction between statisticians and subject area scientists

Statistics 210

Broad outline

- Data collection and randomization
- Comparative (two-population) studies
- Analysis of variance (> 2 samples)
- Blocking/pairing to reduce variance
- Simple linear regression
- Multiple regression
- Analysis of covariance (relation of regression and ANOVA)
- Factorial experiments (return to ANOVA)
- and maybe Nested designs/random effects

Stat 210

Prerequisites

- Calculus and linear algebra
- Introductory or basic statistics class
(descriptive statistics, elementary probability, basic inference for means and proportions)
- Undergraduate probability / statistics
(random variables, probability distributions, moments, joint distributions, basic inference theory)
- References for review of basic statistics (Stat 7):
 - **The Basic Practice of Statistics** - Moore
 - **Statistics** - Freedman, Pisani, Purves, Adhikari
- References for review of undergrad prob/stat (Stat 120ABC):
 - **Mathematical Statistics and Data Analysis** - Rice
 - **An Introduction to Mathematical Statistics and its Applications** - Larsen, Marx

Data collection

- Experiment vs observational study
 - Experiment - investigator intervention to determine the level of one or more factor for each experimental unit (e.g., drug A or B)
 - Observational study - determination of factor levels is outside investigator's control (e.g., smoking)
 - Experiment with randomized assignment of factor levels to experimental units has the potential to establish cause and effect
 - Experiments with nonrandom assignment or observational studies have a difficult time establishing cause and effect (why? confounding)
 - Observational studies are still useful (study diff't groups, establish association as step on causal chain)

Data collection

- Selection of experimental units from the population
 - Random sample from population - can draw inferences for population from sample
 - Nonrandom sample - no inference to population (ex.: 1936 Literary Digest Poll)
 - Simple random sample of size n : every subset of size n in the population has the same probability of being chosen
 - We assume simple random samples
 - Other viable approaches exist for obtaining a representative sample

Data collection

- Idealized story
 - define population of interest
 - random sample of units for study
 - random assignment of factor levels to experimental units
 - analyze data (making assumptions about the population as needed)
 - statistical inference about population
- Common compromises
 - observational study for ethical or other reasons (e.g., smoking)
 - samples of convenience (e.g., Psych 100 students)
 - these compromises are OK, as long as limitations are realized

Comparative studies (two samples)

Experiments

- Experiment - key feature is investigator intervention to change the level of one or more factor keeping others fixed
- Cause/effect relationships can be determined **if experiment is done well**
- Terminology
 - experimental unit - unit to which treatment is applied (person, animal, student/class)
 - factor - quantity/item thought to affect outcome (diff't levels)
 - treatment - combination of levels of factors
 - response/outcome

Comparative studies (two samples)

Experiments

- Principles of experimentation
 - control - eliminate factors not under study (e.g., use control group)
 - randomization - random assignment of treatment to experimental units
 - replication - repeat on many units
 - blocking/matching/pairing - divide experimental units into homogeneous groups and apply each treatment in each group
 - first three of these are needed for causal inference

Comparative studies (two samples)

Experiments

- Example (Salk Polio Vaccine)
 - 1954 field trial
 - two main designs: NFIP (this was not randomized) and a randomized design
 - NFIP: grades 1,3 = control,
grade 2 with consent = vaccine,
grade 2 without consent = control
 - problems with NFIP
 - * grades 1 and 3 not valid controls
(polio contagious)
 - * grade 2 (no consent) not valid control
(diff't types of people)
 - randomized: ask for consent from everyone
consent = randomly assign 1/2 to vaccine
and 1/2 to control
no consent = not in study
 - other control - double blind (controls receive placebo, doctors don't know assignment)

Comparative studies (two samples)

Experiments

- Example (Salk Polio Vaccine) - results

Randomized

group	sample size	rate (per 100K)
Treat	200K	28
Control	200K	71
No consent	350K	46

NFIP

group	sample size	rate (per 100K)
Grade 2 (vacc)	225K	25
Grade 1,3 (control)	725K	54
Grade 2 (no cons)	125K	44

- lines 1 and 3 are similar (as expected)
- line 2's are diff't (NFIP includes some people that would not have consented)

Comparative studies (two samples)

Experiments

- Example (Salk trial) - role of randomization
 - Randomization actually plays two roles in statistical studies
 - * remove effect of confounding factors
 - * provides a basis for inference
 - How does randomization provide a basis for inference?
 - * 56 children got polio in trt group,
142 children got polio in control group
 - * if the vaccine has no effect, then children will have same outcome in either group (i.e., same 198 children would get polio)
 - * because we randomized we can easily calculate the probability that 142 of these 198 would show up in one group by chance (1 in a billion)
 - * this “proves” vaccine is effective

Comparative studies (two samples)

Observational studies

- Observational studies
 - observe the population (no intervention)
 - rely on sample of population
(random sample is best)
 - compare existing groups
 - causal inference is difficult
- Observational studies example 1:
Nurses Health Study
 - 10,000 nurses (female age 20-30 at start)
 - food intake diaries
 - find association between fat and heart disease
 - no control - genetics, exercise, weight, stress
 - useful info but not cause and effect

Comparative studies (two samples)

Observational studies

- Observational studies example 2:
Baltimore Housing Study (1950s)
 - 400 families in housing project (2 deaths)
 - 600 families in slums (10 deaths)
 - all 1000 had applied for housing project;
the 400 were “selected” in some way
 - difficult to reach any conclusion

Comparative studies (two samples)

Notation for two sample studies

- Scenarios:
experiment with two treatments,
observational study with two diff populations
- Notation
 - sample $Y_{11}, Y_{12}, \dots, Y_{1n_1}$ of size n_1 from population 1
 - summary statistics

$$\bar{Y}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} Y_{1j} \quad S_1 = \sqrt{\frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_1)^2}$$

- population parameters for population 1
(mean μ_1 , variance σ_1^2)
- corresponding notation for population 2:
sample $Y_{21}, Y_{22}, \dots, Y_{2n_2}$ of size n_2
summary statistics: \bar{Y}_2 and S_2
population parameters: μ_2 and σ_2^2

Comparative studies (two samples)

Inference questions

- Basic inference questions we address
 - point estimation (estimates for μ_1, μ_2 , $\mu_1 - \mu_2$, etc.)
 - interval estimation (confidence intervals for $\mu_1 - \mu_2$ and other quantities)
 - tests of hypotheses (does $\mu_1 = \mu_2$?)
- Also ask:
 - are assumptions valid
 - effects on inference if assumptions are not valid
 - methods for addressing failed assumptions

Comparative studies (two samples)

Randomization inference

- Assume randomized experiment
- Why randomize?
 - reduce/eliminate bias (recall polio example)
 - creates a probability distn
(under hypothesis of no difference)
- Randomization inference
 - under null hypothesis no treatment effect
 - each unit has same response in either group
 - observed difference is $\bar{y}_1 - \bar{y}_2$
 - is observed difference large?
 - compare to other randomizations
 - very powerful approach - no modeling assumptions
 - confidence interval is possible but hard
- Permutation test - same idea when treatment is not randomized

Comparative studies (two samples)

Two sample example - writing study

- experimental units = writers
- treatments = questionnaire before assignment
(questions emphasize intrinsic or extrinsic rewards)
- random assignment (24 intrinsic, 23 extrinsic)
- response = avg of 12 ratings on 40 pt scale
- data

intrinsic:	12.0	12.0	12.9	13.6	16.6	17.2
	17.5	18.2	19.1	19.3	19.8	20.3
	20.5	20.6	21.3	21.6	22.1	22.2
	22.6	23.1	24.0	24.3	26.7	29.7
extrinsic:	5.0	5.4	6.1	10.9	11.8	12.0
	12.3	14.8	15.0	16.8	17.2	17.2
	17.4	17.5	18.5	18.7	18.7	19.2
	19.5	20.7	21.2	22.1	24.0	

Comparative studies (two samples)

Two sample example - writing study (cont'd)

- data summaries:
 $\bar{y}_1 = 19.88, s_1 = 4.44, \bar{y}_2 = 15.74, s_2 = 5.25$
- data display (histograms, boxplots, stem and leaf plots)
- observed difference in means is 4.14
- 1000 randomizations under null hypothesis
 - 5/1000 randomizations have values > 4.14
(or less than -4.14)
 - extremely unlikely to see a difference this big by chance (two-tailed p-value = .005)
- conclusions
 - questionnaire on intrinsic rewards leads to more creative writing
 - no random sample ... can't necessarily infer that this is true on a bigger population
 - hope sample is representative of bigger population

Comparative studies (two samples)

Model-based inference

- Suppose we assume:
 - Y_{11}, \dots, Y_{1n_1} iid $N(\mu_1, \sigma^2)$
 - Y_{21}, \dots, Y_{2n_2} iid $N(\mu_2, \sigma^2)$
 - two samples are independent
 - note that we are assuming constant variance
- Equivalent to linear model

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

with $\epsilon_{ij} \sim N(0, \sigma^2)$ (iid)

Comparative studies (two samples)

Model-based inference (cont'd)

- Some results
 - $\bar{Y}_1 - \bar{Y}_2$ is an estimator for $\mu_1 - \mu_2$
(unbiased, best linear unbiased)
 - $\bar{y}_1 - \bar{y}_2$ is an estimate for $\mu_1 - \mu_2$
 - $\text{Var}(\bar{Y}_1 - \bar{Y}_2) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$
 - $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ is a pooled estimator for common variance σ^2
- Key result: under assumptions

$$t = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

where $t_{n_1+n_2-2}$ refers to Student's t -distn with $n_1 + n_2 - 2$ d.f.

- t is symmetric like the normal distn
- t has longer tails
- t_∞ is normal distn

Comparative studies (two samples)

Model-based inference (cont'd)

- 100(1 - α)% confidence interval for $\mu_1 - \mu_2$
(follows from t -distn result on previous slide)

$$\bar{Y}_1 - \bar{Y}_2 \pm t_{n_1+n_2-2, 1-\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- classical frequentist interpretation
 - μ_1, μ_2 are fixed unknowns
 - $\bar{Y}_1, \bar{Y}_2, S_p$ are random variables
 - confidence interval is a “random interval”
 - in repeated samples, 95% of such intervals contain the true value (a procedure with good long-term frequency properties)
 - strictly speaking can't say 95% probability for one particular interval (but this casual interpretation is generally OK)
 - note: width depends on d.f., confidence level, σ , sample size

Comparative studies (two samples)

Model-based inference (cont'd)

- Tests of hypotheses
 - null hypothesis $H_o : \mu_1 = \mu_2$ (no difference)
 - alternative hypothesis $H_a : \mu_1 \neq \mu_2$ (two-sided)
or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$ (one-sided)
 - test statistic $t = (\bar{Y}_1 - \bar{Y}_2) / \left(S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$
 - P -value = probability of obtaining a value of the test statistic as big or bigger than the observed value if H_o is true
 - * small P -value means either:
 - (i) H_o is true and we were very unlucky
 - OR (ii) H_o is false
 - * can think of as a measure of evidence
 - * people often use .05 as a formal cutoff
(BAD IDEA)
 - * P -value is NOT the probability that H_o is true
 - * P -value says nothing directly about the alternative (calculated assuming H_o is true!)

Comparative studies (two samples)

Model-based inference (cont'd)

- Two views of testing
 - Hypothesis testing/decision procedures
 - * Neyman-Pearson approach
 - * fix $\alpha = \Pr(\text{reject } H_o \text{ when true})$
 - * develop rejection region (e.g.,
 $|(\bar{x}_1 - \bar{x}_2)/s_p \sqrt{1/n_1 + 1/n_2}| > t_{n_1+n_2-2, 1-\alpha/2}^*$)
 - * conclusion is reject or not (no P -value!!)
 - Significance testing
 - * Fisher
 - * calculate test statistic
 - * compute P -value
 - * report P -value as evidence against null

Comparative studies (two samples)

Model-based inference (cont'd)

- Relationship of tests and confidence intervals
 - powerful idea: duality between tests and CIs
 - test (two-sided) will reject at .05 level (P-value less than .05) if and only if 0 is not in 95% confidence interval
 - can be used, for example, to get randomization confidence intervals
 - * consider possible value for $\delta = \mu_1 - \mu_2$
 - * subtract δ from every value in sample 1
 - * perform randomization test to determine if $Y_{1j} - \delta$'s have different mean than Y_{2j} 's
 - * if don't reject H_o put δ in CI

Comparative studies (two samples)

Writing study example (cont'd)

- Model-based inference
 - 95% CI:
$$4.14 \pm (2.014)(4.85)\left(\sqrt{\frac{1}{24} + \frac{1}{23}}\right) = (1.29, 6.99)$$
 - t -test: $t = \frac{4.14}{(4.85)\left(\sqrt{\frac{1}{24} + \frac{1}{23}}\right)} = 2.93$
 P -value = .005(2 – sided) or .025(1 – sided)
- Randomization vs model-based inference
 - probability by randomization vs probability by assumed population model
 - randomization requires no model for population
 - randomization tests require more computation; randomization CI's (based on test/interval relationship) require even more
 - model-based test and CI are easy (an approximation to randomization inference)
 - model-based approach provides insight into study design (next slide)

Comparative studies (two samples)

Study design

- Sample size for confidence intervals
 - half-width of CI (assuming $n_1 = n_2 = n$) is $t_{2(n-1), 1-\alpha/2} S_p \sqrt{2/n}$
 - can find n to achieve specified half-width
 - one difficulty is that n enters twice (d.f. and sample size)
 - one idea: compute initial guess n_o using $z_{1-\alpha/2}$ in place of t critical value, and then improve guess using $t_{2(n_o-1), 1-\alpha/2}$

Comparative studies (two samples)

Study design (cont'd)

- Power/Sample size for test
 - type I error = reject H_o when it is true
 - type II error = fail to reject H_o when it is false
 - let $\alpha = \text{Pr}(\text{type I error})$
 - let $\beta = \text{Pr}(\text{type II error})$
 - α is known as the size, level, significance
 - $1 - \beta$ is known as the power (probability of correct rejection of H_o)
 - α (size, level, significance) in advance at .05
 - for fixed α : power (or β) is determined by true effect size $\delta = \mu_1 - \mu_2$, true variance σ^2 , and sample sizes
 - key formula (two-sided test w/ equal sample sizes)

$$n = 2(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma^2 / \delta^2$$

- two ideas
 - * given sample size can find β for different δ 's (power function)
 - * determine sample size to achieve a desired β
- Table B.5 gives power for t-test BUT uses $\delta = (\mu_1 - \mu_2) / (\sigma \sqrt{2/n})$

Comparative studies (two samples)

Model diagnostics

- Model-based approach makes a number of assumptions
 - indep samples from each population
 - two populations are independent
 - equal variances
 - normal distribution
- Plan of attack
 - diagnose whether assumption is valid or not (graphical and statistical tools)
 - understand the effects of violated assumptions
 - remedies (modify/transform data or revisit model)

Comparative studies (two samples)

Model diagnostics - independence

- Diagnosis
 - usually design study to achieve independence
 - could fail if units are related
(students in same class)
 - check by looking at residuals ($Y_{ij} - \bar{Y}_i$) within possible clusters
 - check by looking at residuals vs possibly relevant variables like time
- Effects
 - $\text{Var}(\bar{Y}_1 - \bar{Y}_2) \neq \sigma^2(1/n_1 + 1/n_2)$
 - t procedures are in trouble
- Remedies
 - if clustering - reanalyze using the correct unit
 - if time effects - need new time-series models

Comparative studies (two samples)

Model diagnostics - equal variances

- Diagnosis
 - histogram of residuals in the two samples
(not a powerful diagnostic tool)
 - test for equality of variance (F-test)
 - * $F = S_1^2/S_2^2$ has F -distn with $n_1 - 1, n_2 - 1$ d.f.
(Table B.4)
 - * BUT F-test is very sensitive to normal assumption hence not recommended
 - Levene's test - t -test comparing absolute deviations in the two groups
(pp 112-114 in text)
 - rule of thumb - OK if variances are within a factor of two
- Effects of unequal variances
 - minor if sample sizes are the same
 - can be important if $n_1 \neq n_2$ (worst case small sample size with larger variance)
 - is a hypothesis about the means relevant here?

Comparative studies (two samples)

Model diagnostics - equal variances

- Remedies for unequal variances
 - transformation
 - * replace data Y with $Z = g(Y)$
 - * perform inference on Z 's
 - * choosing the transformation
 - trial and error - $\ln Y$ or \sqrt{Y}
 - Box-Cox family of transformations
 Y^λ ($\lambda \neq 0$) and $\ln Y$ ($\lambda = 0$)
 - optimize within family (maximum likelihood estimation)
 - transformation based on science
(sqrt of area, cube root of volume)
 - * interpretation of results can be harder (mean of logarithms is not the logarithm of the mean)
 - * questions: should we account for transformation? are we snooping through the data to find significance?

Comparative studies (two samples)

Model diagnostics - equal variances

- Remedies for unequal variances (cont'd)
 - approximate t -distn
 - * use separate sample variances for the two samples, then

$$t^* = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \text{ is approx } t_\nu$$

where

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{S_2^2}{n_2}\right)^2}$$

is the Cochran-Satterthwaite approximation
with $\min(n_1 - 1, n_2 - 1) \leq \nu \leq n_1 + n_2 - 2$

- * tests or CI based on this approximation

Comparative studies (two samples)

Model diagnostics - normality

- Diagnosis
 - histogram of residuals (pool the residuals from the two samples if we believe equal variance; consider separately if not)
 - normal probability plot of residuals
 - * order data (residuals here) from smallest to largest (say $X_{(1)}, \dots, X_{(n)}$)
 - * find what we would expect if normal (from tables or approximation)
 - * Blom approximation $q_i = \Phi^{-1} \left(\frac{i-.375}{n+.25} \right)$
 - * scatterplot of $X_{(i)}$ vs q_i is straight line if data are normal
 - * curves indicate non-normal tails
 - skewness = $E(Y - \mu)^3 / \sigma^3$ (zero for normal)
 - kurtosis = $E(Y - \mu)^4 / \sigma^4$ (three for normal)
excess kurtosis = kurtosis - 3 (e.g., in SAS)

Comparative studies (two samples)

Model diagnostics - normality (cont'd)

- Diagnosis
 - statistical tests
 - * many exist (Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling, skewness, kurtosis)
 - * Table B.6 in text gives critical values for correlation of normal probability plot (similar to Shapiro-Wilk)
- Effects of non-normal data
 - large samples - no problem because of CLT
 - sensitive to outliers (not resistant)
 - problem if two distn have different shapes
 - if two distn have same shape and equal sample sizes, then skewness is not a problem
 - if two distn have same shape and unequal sample sizes, then skewness can be a problem

Comparative studies (two samples)

Model diagnostics - normality (cont'd)

- Remedies for non-normal data
 - examine outliers
 - * analyze data with and without to see if conclusions change
 - * remove only if one can argue that observations are from a diff't population
 - transformation
 - * discussed earlier (under unequal variances)

Comparative studies (two samples)

Model diagnostics - normality (cont'd)

- Remedies for non-normal data
 - Nonparametric tests (Wilcoxon rank sum)
 - * combine two samples and put in order (smallest to largest)
 - * replace each observation by its rank in combined sample (1=smallest, $n_1 + n_2$ =largest)
 - * test statistic is sum of ranks in sample 1
 - * test statistic is approx normal with
mean = $n_1(n_1 + n_2 + 1)/2$ and
variance = $n_1n_2(n_1 + n_2 + 1)/12$
 - * does not assume normality, equal variance
 - * basically a test on median not mean
 - * very effective test (hard to get CI)

One-way ANOVA (> 2 samples)

Introduction

- Independent random samples from r populations
- Examples
 - randomized experiment with r treatments
 - observational study with r different groups
- Two sources of variation in measurements
 - variability among observations in a group
 - variability among groups
- Question: are differences among groups large relative to variation within groups
- Randomization inference is possible but we won't discuss it
- Model-based inference (called ANOVA model)

One-way ANOVA (> 2 samples)

Notation

- $Y_{ij} = j^{th}$ observation in i^{th} sample
- $i = 1, \dots, r; j = 1, \dots, n_i$
- Data summaries:
 - for i^{th} sample:
 - * sample size n_i
 - * sample mean $\bar{Y}_i = \bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$
 - * sample standard deviation
$$S_i = \sqrt{\frac{1}{n_i - 1} \sum_j (Y_{ij} - \bar{Y}_i)^2}$$
 - total sample size $N = \sum_i n_i$
 - overall mean $\bar{Y} = \bar{Y}_{..} = \frac{1}{N} \sum_i \sum_j Y_{ij}$
 - pooled variance estimate:
$$S_p^2 = \frac{1}{N - r} (\sum_i (n_i - 1) S_i^2)$$

One-way ANOVA (> 2 samples)

Model

- Assume $Y_{ij} \sim N(\mu_i, \sigma^2)$
- Equivalent to $Y_{ij} = \mu_i + \epsilon_{ij}$ with $\epsilon_{ij} \sim N(0, \sigma^2)$
(known as the cell means model)
- Observations independent within group
- Observations independent across groups
- Alternative version of model
 - $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ with
either $\sum \alpha_i = 0$ or $\sum n_i \alpha_i = 0$
 - known as factor effects version
 - α_i measures difference between i th group mean and overall mean (effect)

One-way ANOVA (> 2 samples)

ANOVA table

- Variation and sums of squares
 - $SS_{total} = \sum_i \sum_j (Y_{ij} - \bar{Y})^2 = \sum_i \sum_j Y_{ij}^2 - N\bar{Y}^2$
(sometimes known as total, corrected for overall mean)
 - $SS_{between} = \sum_i \sum_j (\bar{Y}_i - \bar{Y})^2 = \sum_i n_i (\bar{Y}_i - \bar{Y})^2$
(variation among group means)
 - $SS_{within} = \sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2 = \sum_i (n_i - 1) S_i^2$
(total within group variation, sometimes called SS_{error})
 - let $e_{ij} = Y_{ij} - \bar{Y}_i$ denote the residual for the ij^{th} observation (difference between observed and estimated/fitted value)
 - note that $SS_{within} = \sum_i \sum_j e_{ij}^2$
 - key result: $SS_{total} = SS_{between} + SS_{within}$

One-way ANOVA (> 2 samples)
ANOVA table (cont'd)

- Sums of squares are recorded in ANOVA table

source of variation	degrees of freedom	sums of squares	mean square
between groups	$r - 1$	$SS_{between}$	SS/df
within group	$N - r$	SS_{within}	SS/df
total	$N - 1$	SS_{total}	

- Note that MS are computed as corresponding SS divided by appropriate degrees of freedom (df)
- MS_{within} is sometimes known as MS_{error} or MSE
- Under the model:
 - $E(MS_{within}) = \sigma^2$
 - $E(MS_{between}) = \sigma^2 + \frac{1}{r-1} \sum_{i=1}^r n_i (\mu_i - \bar{\mu})^2$
where $\bar{\mu} = \sum_i n_i \mu_i / \sum_i n_i$
 - $F = MS_{between} / MS_{within}$ will be near one if all μ_i 's are equal and larger than one when they are not

One-way ANOVA (> 2 samples)

ANOVA table (cont'd)

- Testing the null hypothesis $H_o : \mu_1 = \mu_2 = \dots = \mu_r$
Under the null:
 - $MS_{within} \sim \sigma^2 \chi_{N-r}^2 / (N - r)$
 - $MS_{between} \sim \sigma^2 \chi_{r-1}^2 / (r - 1)$
 - $MS_{between}$ and MS_{within} are independent
 - $F = MS_{between} / MS_{within}$ has the central F -distn with $r - 1$ and $n - r$ degrees of freedom
 - P -values from Table B.4
- Key point: Rejecting the null is not the end of the analysis. It is not interesting to say we found a model that doesn't fit. We take a brief digression and then return.

One-way ANOVA (> 2 samples)

Fixed effects and random effects

- Fixed effects
 - r treatments are of direct interest
 - only treatments under consideration
 - e.g., two drugs, four pesticides
- Random effects
 - r groups are just a sample from population
 - real questions are about population
 - same model with the additional assumption that $\mu_i \sim N(\mu, \sigma_\mu^2)$
 - don't estimate μ_i , estimate μ and σ_μ^2
 - intraclass correlation = $\sigma_\mu^2 / (\sigma_\mu^2 + \sigma^2)$

One-way ANOVA (> 2 samples)

Fixed effects and random effects (cont'd)

- Random effects example
 - sample 8 AP (high school) statistics classes (these are the groups/treatments)
 - sample 10 students in each class
 - give intro stat course final exam to each
 - not just interested in these 8 classes
 - μ measures overall effectiveness of AP statistics classes
 - σ_{μ}^2 measures variability among AP classes
 - intraclass correlation measures variability among classes relative to variation among students
- For now we used fixed effects; will make occasional comments about random effects and return to the topic late in the quarter (??)

One-way ANOVA (> 2 samples)

Example - Cash offers for cars (problem 16.10)

- Three groups of sellers (young, middle-aged, old)
- 12 individuals in each group
- Each tried to sell same used car
- Response is price offered by dealer (thousands)
- Observational study (why?)

- Data summaries

group	n_i	\bar{y}_i	s_i
young	12	21.50	1.73
middle	12	27.75	1.29
old	12	21.42	1.68

- ANOVA table

source	df	SS	MS
age	2	316.72	158.36
error	33	82.17	2.49
total	35	398.89	

- $F = 158.36/2.49 = 63.6$
- P -value $< .001$ (way less) – reject H_o
- Now what? Which groups are different?
(obvious answer here)

One-way ANOVA (> 2 samples)

Comparisons and contrasts

- Assume we reject hypothesis of equal means
- Types of questions/comparison
 - inference for a single group mean
 - pairwise comparisons
 - linear combinations (contrasts as a special case)
- Inference for a single group mean
 - $100(1 - \alpha)\%$ CI: $\bar{Y}_i \pm t_{N-r, 1-\alpha/2} \sqrt{\frac{MS_{error}}{n_i}}$
 - usual analysis with pooled variance estimate and extra degrees of freedom

One-way ANOVA (> 2 samples)
Comparisons and contrasts (cont'd)

- Pairwise comparisons
 - compare two means using usual two-sample procedures with the common variance σ^2 estimated by MS_{error}
 - 100(1 - α)% CI:
$$\bar{Y}_i - \bar{Y}_j \pm t_{N-r, 1-\alpha/2} \sqrt{MS_{error} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$
 - test that corresponds to the above interval
 - with all n_i equal the quantity
 $t_{N-r, 1-\alpha/2} \sqrt{MS_{error} \left(\frac{2}{n} \right)}$ is known as the LSD (least signif. difference)
 - often see means listed in order with underlining used to indicate “similar” means (by LSD)

One-way ANOVA (> 2 samples)
Comparisons and contrasts (cont'd)

- A possible problem
 - each pairwise comparison has type I error level α or confidence level $100(1 - \alpha)$
 - we do $\binom{r}{2}$ such comparisons!
 - if r is large some significant differences are expected by chance even if all of the means are the same
 - known as the multiple comparisons problem
 - more to come on this subject

One-way ANOVA (> 2 samples)

Comparisons and contrasts (cont'd)

- Linear combinations/contrasts
 - interested in a linear combination $\gamma = \sum_i c_i \mu_i$
 - **contrasts** are the special case where $\sum c_i = 0$ (then γ is zero if all μ_i are equal)
 - pairwise comparisons are also a special case ($c_i = 1, c_j = -1$, and other c 's are zero)
 - point estimate: $\hat{\gamma} = \sum_i c_i \bar{Y}_i$
 - standard deviation: $sd(\hat{\gamma}) = \sigma \sqrt{\sum_i (c_i^2/n_i)}$
 - standard error: $s.e.(\hat{\gamma}) = \sqrt{MS_{error} \sum_i (c_i^2/n_i)}$
 - 100(1 - α)%CI: $\hat{\gamma} \pm t_{N-r, 1-\alpha/2} s.e.(\hat{\gamma})$
 - test $H_o : \gamma = 0$ by computing $t = \hat{\gamma}/s.e.(\hat{\gamma})$ (compare to t -distn with $n-r$ d.f.)
 - above test is of great interest if c is a contrast
 - contrasts are 1 d.f. tests of (usually prespecified) hypotheses
 - contrasts can be used to decompose the $SS_{between}$
 - sum of squares for contrast c is $SS_c = \hat{\gamma}^2 / (\sum_i c_i^2/n_i)$

One-way ANOVA (> 2 samples) Comparisons and contrasts (cont'd)

- Where do contrasts come from?

(illustrate assuming $r = 5$)

– comparing subgroups

* pairwise comparison, e.g., $c = (1, -1, 0, 0, 0)$

* groups 1,2,3 vs groups 4,5: $c = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, -\frac{1}{2}, -\frac{1}{2})$

$$\gamma = \frac{\mu_1 + \mu_2 + \mu_3}{3} - \frac{\mu_4 + \mu_5}{2}$$

– expected trends

* linear trend in means: $c = (-2, -1, 0, 1, 2)$

* quadratic trend in means: $c = (-2, 1, 2, 1, -2)$

* orthogonal polynomials (like above) can be used to analyze ordered treatments

– any hypothesis can become a contrast

* if we expect means

$\mu_1 = 3, \mu_2 = \mu_3 = 7, \mu_4 = \mu_5 = 4$, then subtracting the overall mean (5) yields weights

$$c = (-2, 2, 2, -1, -1)$$

One-way ANOVA (> 2 samples)

Comparisons and contrasts (cont'd)

- Orthogonal contrasts
 - two contrasts (given by weights c and b) are orthogonal if $\sum_i b_i c_i / n_i = 0$
 - estimates for such contrasts are uncorrelated
 - $r - 1$ pairwise orthogonal contrasts will completely decompose $SS_{between}$
 - there is more than one possible choice for the $r - 1$ orthogonal contrasts
 - return to example - Cash offers
 - * 3 age groups have natural offering
 - * natural to decompose $SS_{between}$ into linear and quadratic contributions
 - * linear ($c = (-1, 0, 1)$) $SS_c = .04$
don't reject $H_o : \sum_i c_i \mu_i = 0$
 - * quadratic ($c = (-1, 2, -1)$) $SS_c = 316.68$
reject H_o
 - * means are not equal and appear to follow a quadratic pattern (obvious from graph)

One-way ANOVA (> 2 samples)

Multiple comparisons

- Typical study has a relatively small number of planned comparisons or contrasts
- Often, investigators consider a large number of unplanned comparisons (e.g., all pairwise comparisons among treatments)
- Problem is that there can be many such “unplanned” comparisons
- Need to adjust tests so that experiment-wide type I error rate is reasonably small to avoid “falsely” significant findings
- What to do?
- Basic approach is to adjust the $t_{N-r, 1-\alpha/2}$ quantity used in $100(1 - \alpha)\%$ confidence intervals and tests of size α

One-way ANOVA (> 2 samples)

Multiple comparison procedures

- Bonferroni
 - if we have m tests/intervals, then use α/m instead of α in each test/interval
 - conservative
(experiment-wide type I error rate $< \alpha$)
 - easy to do
 - need to know number of comparisons m
- Scheffe
 - works for any number of (actually all possible) tests/intervals
 - most conservative, still relatively easy
 - use $\sqrt{(r-1)F_{r-1, N-r, 1-\alpha}}$ in place of $t_{n-r, 1-\alpha/2}$
- F-protected LSD (weak)
 - do overall F-test first
 - if reject H_o then use LSD for pairwise comparisons differences)

One-way ANOVA (> 2 samples)

Multiple comparison procedures (cont'd)

- Tukey-Kramer (studentized range)
 - an exact solution for all pairwise comparisons
 - experiment-wide error rate is α over all of the $\binom{r}{2}$ possible comparisons
 - assumes equal sample size = n
(conservative if not)
 - distn of $q(r, N - r) = (\max_i \bar{Y}_i - \min_i \bar{Y}_i) / (S_p / \sqrt{n})$ is in Table B.9 (this is the most extreme pairwise comparison)
 - use $\frac{1}{\sqrt{2}}q(r, N - r, 1 - \alpha)$ in CIs
 - for tests, take $\sqrt{2}$ times usual t -test statistics and compare to q distn

One-way ANOVA (> 2 samples)

Multiple comparison procedures - a different look

- False discovery rate
 - previous approaches attempt to control experimentwise error rate
 - most appropriate if serious concern that all null hypotheses are true
 - alternative is to control percentage of discoveries (statistically significant results) that are false
 - FDR (Benjamini and Hochberg, JRSS B, 1995)
 - * order P -values from smallest (most significant) to largest
 - * find $k =$ largest index such that $P_{(i)} \leq i * q/m$ (where q is desired FDR and m is number of tests)
 - * reject hypotheses corresponding to smallest k P -values
 - Problem: FDR as above assumes independent tests
 - * later publication suggests replacing q by $q / \sum_{j=1}^n 1/j$ to correct for non-independence

One-way ANOVA (> 2 samples)

Model diagnostics

- Model assumptions
 - indep samples from each population
 - populations are independent
 - equal variances
 - normal distributions

One-way ANOVA (> 2 samples)

Model diagnostics - independence

- Similar to two sample case
- Diagnosis
 - usually design study to achieve independence
 - could fail if units are related (students in same class)
 - check by looking at residuals ($Y_{ij} - \bar{Y}_i$) within possible clusters
 - check by looking at residuals vs possibly relevant variables like time
- Effects of non-independence
 - wrong variances for means
 - wrong error terms for tests
- Remedies for non-independence
 - include the source of the correlation in the model

One-way ANOVA (> 2 samples)

Model diagnostics - equal variances

- Diagnosis
 - histogram of residuals in the samples
 - test for equality of variance (Bartlett's test)
 - * $M = (\sum_{i=1}^r (n_i - 1)) \log S_p^2 - \sum_{i=1}^r (n_i - 1) \log S_i^2$
 - * $C = 1 + \left(\frac{1}{3(r-1)} \right) \left(\sum_i \frac{1}{n_i-1} - \frac{1}{N-r} \right)$
 - * $X^2 = M/C$ is approx χ^2 on $r - 1$ d.f.
 - * other tests (Hartley, Levene)
 - * BUT tests are very sensitive to normal assumption hence not recommended
 - rule of thumb - OK if variances within a factor of four
- Effects of non-constant variance
 - minor if sample sizes are the same
 - can be important if very unequal sample sizes
 - is a hypothesis about the means relevant here?

One-way ANOVA (> 2 samples)

Model diagnostics - equal variance (cont'd)

- Remedies for nonconstant variance

- transformation

- * replace Y_{ij} by $f(Y_{ij})$

- * rules of thumb:

- data are all positive - use $\log Y_{ij}$

- data are proportions - use $\arcsin \sqrt{Y_{ij}}$

- data are counts - use $\sqrt{Y_{ij}}$ or $\sqrt{Y_{ij} + 1}$

- * if $\text{Var}(Y_{ij}) = g(E(Y_{ij}))$, then variance stabilizing transformation is

$$h(y) \propto \int \frac{1}{\sqrt{g(z)}} dz$$

- * examples:

- if $\text{var} \propto \text{mean}$, then $g(z) = z$ and $h(y) = \sqrt{y}$

- if $\text{var} \propto \text{mean}^2$, then $g(z) = z^2$ and $h(y) = \ln y$

- * with many groups can estimate $g(z)$

- if $\text{var} \propto \text{mean}^{2\lambda}$, then λ is slope in plot of $\log S_i$ vs $\log \bar{Y}_i$

- compute $\log S_i$ and $\log \bar{Y}_i$ for each group

- estimate λ from scatterplot

- transform is $h(y) = y^{1-\lambda}$ (or $\ln y$ if $\lambda = 1$)

One-way ANOVA (> 2 samples)

Model diagnostics - equal variance (cont'd)

- Remedies for nonconstant variance
 - weighted least squares
 - * easiest to describe in regression context
 - * basic idea is to weight observations in each group according to $1/S_i^2$
 - * more later
 - nonparametrics - Kruskal-Wallis test
 - * combine all groups
 - * rank the observations
 - * ANOVA on the ranks
 - * $KW = SS_{between}(N - 1)/SS_{total}$
 - * compare KW to χ_{r-1}^2 distn
 - * related to usual F test on ranks
 - * pairwise comparisons are possible
(generalizes two-sample rank sum procedure)

One-way ANOVA (> 2 samples)

Model diagnostics - normality

- Diagnosis
 - normal probability plot of residuals
(pooled or separate in each group)
 - statistical tests
(see two-sample notes for details)
- Effects
 - inaccurate inferences if badly non-normal
 - ANOVA tests are sensitive to outliers

One-way ANOVA (> 2 samples)

Model diagnostics - normality (cont'd)

- Remedies for non-normality
 - transformation
 - * see comments under non-constant variance
 - * hope that same transformation achieves both goals
 - * achieving constant variance is more important
 - non-parametrics
 - * Kruskal-Wallis test described above

One-way ANOVA (> 2 samples)

Study design

- Power calculation
 - assume completely randomized design
 - assume r populations, equal sample size n
 - select $\alpha = \Pr(\text{type I error})$
 - key to calculating power is non-centrality parameter of the F -test
(measuring degree to which means differ)

$$\phi = \frac{1}{\sigma} \sqrt{\frac{n}{r} \sum_i (\mu_i - \bar{\mu})^2}$$

- power table B.11 in NKNW gives power for given α, ϕ
(note that $\nu_1 = r - 1, \nu_2 = N - r$)
- example ($r = 4$)
 - * $\mu_1 = 12, \mu_2 = 13, \mu_3 = 18, \mu_4 = 21$
 - * $\sigma = 2.5, n = 5$
 - * $\phi = \frac{1}{2.5} \sqrt{\frac{5}{4} 54} = 3.29$
 - * $\nu_1 = 3, \nu_2 = 15, \alpha = .01$
 - * power = 1.00

One-way ANOVA (> 2 samples)

Study design

- Sample size calculation
 - can get sample size by trial and error
 - * for example on previous slide:
 - $n = 2$ gives power .8
 - $n = 3$ gives power .99
 - sample size tables exist
 - * Table B.12 - uses Δ/σ
 - where $\Delta = (\max \mu_i - \min \mu_i)$
 - as key quantity in place of ϕ
 - * Nelson, Journal of Quality Technology, 1985 (uses ϕ -like quantity)

Pairing/blocking to reduce variance

Introduction

- Key to inference is comparing observed difference to measure of variability (e.g. $t = (\bar{Y}_i - \bar{Y}_j) / s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$)
- If σ (and hence s_p) is large then it is hard to declare differences significant
- σ is a measure of heterogeneity of population
- Idea: create homogeneous groups and compare treatments within the homogeneous groups
- Examples:
 - matched pairs for comparing two treatments
 - blocking in ANOVA

Pairing/blocking to reduce variance

Paired responses

- Example: agricultural experiment in different fields with treatments A and B applied to half of each field
- Example: medical study - find pairs of people with same gender and age, randomly assign treatments within the pair
- Notation
 - Y_{1k} = response to treatment 1 in k th pair
 - Y_{2k} = response to treatment 2 in k th pair
 - $k = 1, \dots, n$
 - responses are not independent, usually positively correlated
 - μ_1 = mean for population under treatment 1
 - μ_2 = mean for population under treatment 2
 - define $D_k = Y_{1k} - Y_{2k}$ (difference)

Pairing/blocking to reduce variance

Paired responses - inference

- One sample inference using differences
- Note $\mu_d = \mu_1 - \mu_2$ (quantity of interest)
- Let $\bar{D} = \frac{1}{n} \sum_i D_i$ and $s_d = \sqrt{\frac{1}{n-1} \sum_i (D_i - \bar{D})^2}$
- $100(1 - \alpha)\%$ CI for $\mu_1 - \mu_2$ is
$$\bar{D} \pm t_{n-1, 1-\alpha/2} s_d / \sqrt{n}$$
- test for $H_o : \mu_1 = \mu_2$ (equivalent to $\mu_d = 0$) is one-sample t -test:
compare $t = \bar{D} / (s_d / \sqrt{n})$ to t distn with $n - 1$ d.f.

Pairing/blocking to reduce variance

Paired responses vs unpaired responses

- To keep things simple suppose $\sigma_1^2 = \sigma_2^2$
- Also suppose we know common σ^2 and the correlation of Y_{1k} and Y_{2k} (call it ρ)
- $\text{Var}(D_k) = \sigma_d^2 = 2\sigma^2(1 - \rho)$
- Paired analysis: $z = \bar{D} / \sqrt{\sigma_d^2/n}$
- Two indep sample analysis: $z = (\bar{Y}_1 - \bar{Y}_2) / \sqrt{2\sigma^2/n}$
- If $\rho = 0$, then same test statistic (but paired analysis has fewer d.f.)
- If $\rho > 0$, then $\sigma_d^2 < 2\sigma^2$ and increased precision will likely more than compensate for loss in d.f.
- More detailed analysis in Snedecor and Cochran (above is simplistic)
- Paired analysis provides no benefit for estimating μ_1 (or μ_2) alone

Pairing/blocking to reduce variance

Paired analysis - model diagnostics

- Model assumptions
 - need pairs independent of each other
 - differences have normal distribution
- Independence
 - usually guaranteed by design
 - big problem if this fails
 - need to improve model

Pairing/blocking to reduce variance

Paired analysis - model diagnostics

- Non-normality
 - diagnosis - probability plot on differences, statistical tests (as before)
 - effects
 - * OK if sample size is large
 - * may produce invalid inference in small samples
 - * sensitive to outliers
 - remedies
 - * transformations may not work well because differences can be negative

Pairing/blocking to reduce variance

Paired analysis - model diagnostics

- Non-normality remedies (cont'd)
 - nonparametric tests
 - * Wilcoxon signed rank test
 - order absolute value of differences
 - assign ranks
 - test statistic = T = sum of ranks corresponding to positive differences
 - T is approximately normal (for large n)
mean $n(n + 1)/4$ and
variance $n(n + 1)(2n + 1)/24$
 - not quite as good as t -test for normal data, but
much better than t for other distn

Pairing/blocking to reduce variance

Paired analysis - model diagnostics

- Non-normality remedies (cont'd)
 - nonparametric tests
 - * sign test
 - ignore zero differences
 - count # of positive differences
 - test hypothesis that proportion of positive differences is 0.5
 - use binomial probability distn
 - ex: suppose 7 of 8 differences are positive
 $P\text{-value} = \Pr(7 \text{ or } 8 \text{ out of } 8 \text{ when } p = 0.5)$
 $= .070$
 - easy, but not as powerful as Wilcoxon

Pairing/blocking to reduce variance

Randomized (complete) block design

- Generalization of the paired response idea
- Assume we have one-way ANOVA with J treatments (new notation)
- Group experimental units into I blocks of size J
- Within each block randomly assign J treatments to units
- Each block is essentially a repetition of experiment
- Remarks
 - if block has too few units for full repetition, then we can use incomplete block design
 - if block has many units (some multiple of J), then we can apply each treatment more than once

Pairing/blocking to reduce variance

Randomized block design

- Model

$$Y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$$

- $i = 1, \dots, I$ indexes blocks
- $j = 1, \dots, J$ indexes treatments
- τ_j are treatment effects (with $\sum_j \tau_j = 0$)
- β_i are block effects (with $\sum_i \beta_i = 0$)
- all effects are assumed to be fixed effects
- additive model
(same treatment effect in each block)

- ANOVA table

source of variation	degrees of freedom	sums of squares
blocks	$I - 1$	$J \sum_{i=1}^I (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$
treatments	$J - 1$	$I \sum_{j=1}^J (\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2$
error	$(I - 1)(J - 1)$	$\sum_{i=1}^I \sum_{j=1}^J (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2$
total	$IJ - 1$	$\sum_{i=1}^I \sum_{j=1}^J (Y_{ij} - \bar{Y}_{\cdot\cdot})^2$

Pairing/blocking to reduce variance

Randomized block design (cont'd)

- Inference
 - usually no test for block effects because blocks are known to be different
 - $E(MS_{treatments}) = \sigma^2 + \frac{I}{J-1} \sum_{j=1}^J \tau_j^2$
 - $E(MS_{error}) = \sigma^2$
 - test for treatment effect

$$F = \frac{MS_{treatments}}{MS_{error}} \text{ is } F_{J-1, (I-1)(J-1)}$$

- assuming F -test is significant:
inference for means, pairwise comparisons,
contrasts, proceeds as in one-way ANOVA

Pairing/blocking to reduce variance

Blocks as random effects

- Often wish to think of blocks as random effects
- Model now assumes $\beta_i \sim N(0, \sigma_\beta^2)$
- Comments made about random effects with one-way design apply here
- Analysis is essentially unchanged except that intraclass correlation may now be of interest

Pairing/blocking to reduce variance

Efficiency of blocking

- In randomized block design $\hat{\sigma}_{rb}^2 = MS_{error}$
- To assess effectiveness of blocking we need to figure out what error variance might have been without blocking
- Snedecor and Cochran give

$$\hat{\sigma}_{cr}^2 = \frac{(I - 1)MS_{blocks} + I(J - 1)MS_{error}}{IJ - 1}$$

as an unbiased estimate of error variance for completely randomized design (proof in Cochran and Cox, 1957)

- Then $\hat{\sigma}_{cr}^2 / \hat{\sigma}_{rb}^2$ is efficiency of blocking (big values mean big benefits)
- One complication is that there are different degrees of freedom associated with these estimates – Fisher measured “information” by multiplying variance estimates by $(df + 3)/(df + 1)$ to adjust for difference in d.f.

Pairing/blocking to reduce variance

Randomized block design - diagnostics

- Assumptions
 - blocks should be independent
 - homogeneous error variance
 - normality
 - block and treatment effects are additive
(no interaction)
- Independence
 - usually arranged by design
 - big problems if violated

Pairing/blocking to reduce variance

Randomized block design - diagnostics

- Constant variance/normality
 - diagnosis
 - * check by examining residuals
$$e_{ij} = Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot}$$
 - * normal probability plot
 - effects
 - * similar to effects for one-way design
 - remedies
 - * transformations
 - * nonparametric test (replace data by ranks within blocks and analyze by ANOVA)

Pairing/blocking to reduce variance

Randomized block design - diagnostics

- Additivity
 - diagnosis
 - * Tukey's test for non-additivity
 - compute

$$SS_{na} = \frac{\left(\sum_i \sum_j (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})Y_{ij}\right)^2}{\sum_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \sum_j (\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2}$$

- compute adjusted SS_{error} as
$$SS_{error.na} = SS_{error} - SS_{na}$$
- F-test of $SS_{na}/MS_{error.na}$ on 1 and $(I - 1)(J - 1) - 1$ d.f.
- effects
 - * more important than normality and constant variance
 - * difficult to interpret treatment effects because effects are different in different blocks
- remedies
 - * try transformations to remove interaction

Pairing/blocking to reduce variance

Randomized block design - study design

- Power and sample size determination for test regarding treatments is as in one-way ANOVA except
 - σ^2 is likely smaller after accounting for blocks
 - error d.f. will be smaller
- To determine number of blocks
 - can pick I to get desired accuracy for specified treatment mean or contrast
 - e.g., $\text{Var}(\bar{Y}_{.j}) = \sigma^2/I$
 - need estimate of σ^2 and can then determine I for needed accuracy

Pairing/blocking to reduce variance

More than one blocking factor

- Can use a broader definition of blocks
- Example: if gender and age are both blocking factors, then one could use as blocks:
males 20-29, males 30-39, females 20-29, etc.
- Problem - this may require many blocks and therefore many experimental units
- Special case: Latin square (two blocking variables with number of levels equal to number of treatments) two blocking factors
 - ex: 3 treatments w/ day and operator as blocking variables

	Operator		
Day	1	2	3
Monday	B	A	C
Tuesday	A	C	B
Wednesday	C	B	A